Ignoring Heterozygous Sites Biases Phylogenomic Estimates of Divergence Times: Implications for the Evolutionary History of *Microtus* Voles

Heidi E.L. Lischer,[1,2] Laurent Excoffier,[1,2] and Gerald Heckel*[,1,2]

[1]Computational and Molecular Population Genetics (CMPG), Institute of Ecology and Evolution, University of Bern, Bern, Switzerland

[2]Swiss Institute of Bioinformatics, Genopode, Lausanne, Switzerland

***Corresponding author:** E-mail: gerald.heckel@iee.unibe.ch.

Associate editor: Beth Shapiro

Raw sequence reads have been deposited in the Sequence Read Archive (SRA) (accession no. SRP034588).

Abstract

Phylogenetic reconstruction of the evolutionary history of closely related organisms may be difficult because of the presence of unsorted lineages and of a relatively high proportion of heterozygous sites that are usually not handled well by phylogenetic programs. Genomic data may provide enough fixed polymorphisms to resolve phylogenetic trees, but the diploid nature of sequence data remains analytically challenging. Here, we performed a phylogenomic reconstruction of the evolutionary history of the common vole (*Microtus arvalis*) with a focus on the influence of heterozygosity on the estimation of intraspecific divergence times. We used genome-wide sequence information from 15 voles distributed across the European range. We provide a novel approach to integrate heterozygous information in existing phylogenetic programs by repeated random haplotype sampling from sequences with multiple unphased heterozygous sites. We evaluated the impact of the use of full, partial, or no heterozygous information for tree reconstructions on divergence time estimates. All results consistently showed four deep and strongly supported evolutionary lineages in the vole data. These lineages undergoing divergence processes split only at the end or after the last glacial maximum based on calibration with radiocarbon-dated paleontological material. However, the incorporation of information from heterozygous sites had a significant impact on absolute and relative branch length estimations. Ignoring heterozygous information led to an overestimation of divergence times between the evolutionary lineages of *M. arvalis*. We conclude that the exclusion of heterozygous sites from evolutionary analyses may cause biased and misleading divergence time estimates in closely related taxa.

***Key words: Microtus arvalis*, rodents, AFLP, high-throughput sequencing, phylogenetics, heterozygosity.**

Introduction

For a long time, the reconstruction of the evolutionary history of closely related species was largely limited to single or a few selected loci. However, a single locus often does not provide enough information to unambiguously infer species relationships, or it can give misleading results due to incomplete lineage sorting, gene flow, gene duplication, recombination, or selection (Maddison 1997; Nichols 2001; Fink et al. 2010). In the last few years, phylogenomic studies aiming at reconstructing evolutionary history based on multiple loci or genomic regions have become more accessible and helped to resolve difficult phylogenetic problems (e.g., Meredith et al. 2011; Struck et al. 2011; von Reumont et al. 2012). Many of these genome-wide analyses have shown significant differences in the evolutionary histories of particular genes and genome regions (Waters et al. 2010; Struck et al. 2011; von Reumont et al. 2012). Inferring phylogenetic trees based on multiple loci can indeed reduce the variance associated with the coalescent process or with sampling errors (Arbogast et al. 2002; Delsuc et al. 2005; Yang and Rannala 2012). Therefore, analyzing sequences from multiple independent loci of a few individuals can result in a more accurate phylogenetic tree than analyzing many individuals at just one genetic locus (Heled and Drummond 2010). Genome-wide analyses of sequence polymorphisms are thus highly desirable but still challenging.

Analytical drawbacks of earlier methods for the reconstruction of phylogenomic trees based on multiple loci have led to the development of more advanced species tree approaches (Yang and Rannala 2012), where species and gene trees are jointly estimated from multi locus data under a hierarchical Bayesian model using coalescent theory (Rannala and Yang 2003; Edwards et al. 2007; Heled and Drummond 2010). These methods can account for deep coalescence (incomplete lineage sorting) or gene tree variability among loci, but cannot deal with migration, gene duplication, or gene loss events (Edwards et al. 2007; Liu and Pearl 2007; Liu 2008; Liu et al. 2010). Species tree approaches are computationally very intensive, and their application to data sets covering very many loci is thus often not feasible (Leache and Rannala 2011; O'Neill et al. 2013), but new methods allow one to overcome some of these restrictions at least

Article

partially (Bryant et al. 2012; Boussau et al. 2013). Other methods estimating species trees based on gene trees under coalescent models are computationally less intensive but do not incorporate gene tree uncertainty (Kubatko et al. 2009; Liu et al. 2009). Distance-based methods (such as neighbor joining [NJ] or UPGMA) on concatenated data have also been suggested as an alternative for large data sets, because they may be able to infer the correct species tree even when gene trees are highly heterogeneous (Liu and Edwards 2009).

Phylogenomic reconstructions within species or between closely related species harbor additional difficulties as they potentially contain a relatively high proportion of heterozygous sites (Sota and Vogler 2003). The analysis of this diploid information is problematic, because heterozygous sites are usually not handled well in phylogenetic methods. One way to overcome this problem is to reconstruct haplotypes from diploid individuals (e.g., with PHASE [Stephens et al. 2001; Stephens and Donnelly 2003] or BEAGLE [Browning SR and Browning BL 2007]). However, phasing distant loci or phasing all heterozygous positions within loci is often not possible (Garrick et al. 2010), and thus, it is not clear how to best join sequences in concatenation-based phylogenomic analyses. As a consequence, heterozygous positions are often partially or totally excluded from further analyses, but their exclusion leads to a reduction of information and might bias phylogenetic parameter estimates.

We study here the influence of different heterozygous single-nucleotide polymorphism (SNP) handling approaches on phylogenomic estimations of the evolutionary history of the common vole (*Microtus arvalis*), a genetically highly polymorphic small mammal species. *Microtus arvalis* is a very abundant rodent distributed over most of Europe where it lives in open habitats such as grass- and farm land from sea level up to around 2,000 m altitude (Hausser 1995; Heckel et al. 2005; Braaker and Heckel 2009; Fischer et al. 2011). Previous phylogenetic analyses of mitochondrial DNA (mtDNA) sequences revealed in Europe four major, geographically distinct evolutionary lineages in the species, which were originally estimated to have diverged clearly before the last glaciation maximum (LGM) around 20,000 years before present (Haynes et al. 2003; Fink et al. 2004; Tougard et al. 2008). A further mtDNA lineage has been described from the southern Balkans (Buzan et al. 2010), and the most divergent lineage occurring further east in Russia is sometimes considered to be a separate species *M. obscurus* (Mitchell-Jones et al. 1999). Such patterns of allopatric divergence in multiple evolutionary lineages appear to be relatively common in the explosive radiation of the *Microtus* genus (Fink et al. 2010). However, it is noteworthy that population-based analyses of nuclear microsatellite and mtDNA data provided much more recent divergence time estimates between the four major European lineages in *M. arvalis* than phylogenetic analyses of mtDNA (Heckel et al. 2005). These discrepancies could be partially due to the maternal inheritance and smaller effective population size of mtDNA, which can lead to a more rapid sorting of lineages compared with nuclear DNA, and due to issues with the calibration of evolutionary rates and other locus-specific effects (Fink et al. 2010).

In this study, we take a genomic approach to investigate the evolutionary history and divergence of the deep evolutionary lineages in *M. arvalis*. The fast development in high-throughput sequencing has made it possible to obtain genome-wide DNA sequences for nonmodel organisms, but sequencing multiple whole genomes often still remains prohibitive especially in organisms for which no genome resources are available (Gaggiotti 2010; Leroux et al. 2010). An alternative and cost-effective approach is the high-throughput sequencing of reduced representation libraries of genomes using approaches such as CRoPS (AFLPseq) or RADseq (van Orsouw et al. 2007; Baird et al. 2008; Leroux et al. 2010; Davey et al. 2011). Ideally, these reduced representation libraries consist of a wide range of independently segregating loci distributed over the whole genome (Horvath et al. 2008). Amplified fragment length polymorphisms (AFLPs) are based on the polymerase chain reaction (PCR) amplification of genome-wide distributed DNA sequences (Vos et al. 1995; Bonin et al. 2006; Fink et al. 2010; Fischer et al. 2011), and high-throughput sequencing of these provides direct access to a large number of loci even for nonmodel species without prior sequence information. AFLPseq has the general advantage that the beginning and end sequences of the fragments are known because these represent restriction enzyme cutting sites, which makes it easier to filter wrong assemblies. This simplified identification of misassemblies is useful because the analysis pipelines are less developed compared with RADseq. However, as an additional difference compared with typical RADseq, longer sequences per locus may enable the capture of multiple polymorphic positions, a feature useful for phylogenomic analyses. We use here data from an AFLPseq experiment to investigate different strategies for handling heterozygous SNPs and study their impact on phylogenomic estimates of tree topology, relative branch lengths, and divergence times, which are the main focus of many evolutionary analyses of closely related organisms.

New Approaches

We developed an approach called repeated random haplotype sampling (RRHS), which allows the integration of information from all alleles at heterozygous sites into phylogenetic tree estimation with, for example, NJ, maximum-likelihood (ML), and Bayesian methods. In the RRHS strategy, haploid sequences for each individual are generated by randomly picking a haplotype from the alleles for each of the sampled loci. A tree is inferred from the random haplotypes, and the process of haplotype generation and tree inference is repeated many times. These trees resulting from RRHS provide comprehensive information on the extent of the variation in phylogenetic results due to allelic and haplotypic variation. These trees can be summarized by calculating a majority rule consensus tree with mean branch length. We provide a Java command line program performing RRHS (http://www.cmpg.iee.unibe.ch/content/softwares__services/computer_programs/rrhs/, last accessed January 9, 2014), which allows

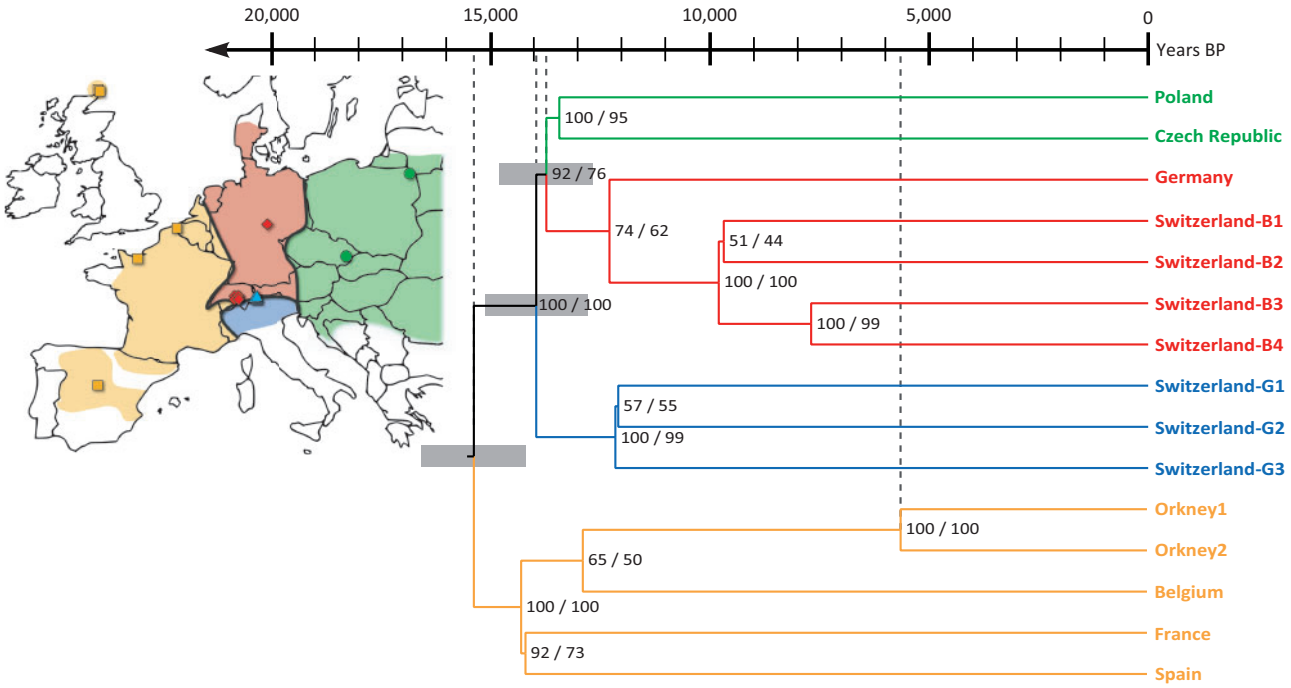

Fig. 1. Phylogenomic tree reveals intraspecific evolutionary lineages. Midpoint-rooted NJ tree based on 1,552 loci in 15 *Microtus arvalis* individuals sampled across Europe (see map). Support values for nodes from 5,000 RRHS trees are given before slashes and bootstrap values after (see Materials and Methods). Divergence times are calibrated on the Orkney individuals' split and are given in years before present (BP). The gray bars show standard deviations of divergence time. The colors on the map and tree represent the evolutionary lineages Eastern (green), Central (red), Italian (blue), and Western (orange).

the integration of this new approach into existing phylogenetic analysis pipelines.

Results

Sequence Polymorphism

The sequencing of AFLP fragments from 15 *M. arvalis* individuals sampled in separate populations across Europe (see fig. 1 and supplementary fig. S1, Supplementary Material online) and the assembly of 1,197,275 sequence reads are summarized in table 1 (see supplementary material, Supplementary Material online, for details; Sequence Read Archive accession SRP034588). In brief, we obtained 38,582 sequence contigs across all individuals after quality filtering, trimming, and assembly of raw sequence reads. A total of 1,741 contigs were shared between all 15 individuals of which 1,552 were polymorphic (total length: 192,623 bp; mean coverage: 259.7). These contained 6,807 SNP positions of which 6,018 (88.4%) were heterozygous in at least one individual. At the individual level, the two Orkney voles contained the lowest numbers of heterozygous SNPs (548 and 576), whereas the two Eastern European individuals held the largest numbers (Czech Republic: 1,263; Poland: 1,131). The average number of nucleotide differences between individuals was lowest between the Orkney individuals (555) and highest between Spain and the rest (fig. 2).

Phylogenomic Analysis

Independent of using our new approach RRHS or the exclusion of heterozygous information (see later), all phylogenomic analyses of the vole data irrespective of the algorithm used (NJ, ML, and Bayesian) inferred the same topological splits separating the 15 individuals in four clusters (fig. 1; supplementary table S1, Supplementary Material online). These four clusters are largely consistent with geography and correspond to the Eastern, Central, Italian, and Western evolutionary lineages in *M. arvalis* inferred from mtDNA (Haynes et al. 2003; Fink et al. 2004; Heckel et al. 2005; Tougard et al. 2008). The genomic separation of individuals according to evolutionary lineages was also supported by principal component analysis (PCA) (supplementary fig. S2, Supplementary Material online).

For estimating the time of divergence between the evolutionary lineages, we used a calibration based on the maximum divergence time between the two individuals sampled in separate populations on the Orkney Islands (supplementary fig. S1, Supplementary Material online). According to paleontological records, *M. arvalis* has never been present in Britain except for the Orkney Islands north of Scotland which became inhabitable only after the Scandinavian ice sheet retreated after the last glaciation (Berry and Rose 1975; Yalden 1982; Corbet 1986; Yalden 1999; Haynes et al. 2003; Martínková et al. 2013). *Microtus arvalis* was introduced in Orkney by Neolithic settlers from the coastal region of Belgium/France 5,000–5,600 years ago (5–5.6 Ka) as determined by recent population genetics analyses and radiocarbon dating of subfossil vole bones (Martínková et al. 2013). Therefore, we used 5.6 Ka as the oldest possible time for divergence between the two Orkney samples.

Table 1. Summary of Roche 454 Sequencing and Assembly of AFLP Loci from 15 *Microtus arvalis* Individuals Sampled across the Species Range.

Reads	Total reads	1,197,275
	Mean of reads per individual	79,818
	Min reads per individual	49,608
	Max reads per individual	107,438
	Average read length	127.8 bp
Contigs	Mean number of contigs per individual	6,599
	Mean coverage	9.4
	Mean number of reads assembled	59,309
	Mean contig length	151.4 bp
	Max contig length	695
	Mean number of singletons per individual	6,754
Assembly between individuals	Number of contigs	38,582
	Polymorphic contigs	11,047
	SNPs	36,408
Shared by all 15 individuals	Contigs	1,741
	Polymorphic contigs	1,552
	Total length of polymorphic contigs	192,623 bp
	Mean coverage	259.7
	SNPs	6,807
	Heterozygous SNP positions	6,018
	Phased SNPs	2,957

Based on this calibration, NJ trees applying 5,000 rounds of RRHS showed divergence times of the Western lineage around 15.3 ± standard deviation 1.2 Ka, the Italian lineage 13.9 ± 1.2 Ka, and the Central from the Eastern lineage 13.7 ± 1.1 Ka (fig. 1; table 2). Divergence time estimates based on individual rounds of random sampling of haplotypes ranged between 11.3 and 20.2 Ka for the Western split, 10.3–18.8 Ka for Italian and 10.3–18.3 Ka for Central and Eastern lineage split. ML and Bayesian algorithms based on RRHS led to slightly older divergence time estimates between 17.0 and 20.4 Ka (for details see supplementary table S1, Supplementary Material online).

The phylogenomic tree inferred with the species tree program SNAPP (Bryant et al. 2012) suggested a somewhat different order of divergence between the evolutionary lineages compared with the other methods (fig. 3). The most likely topology showed the Central and Italian lineages as sister clades, whereas the topology supported by the other methods (Central and Eastern as sister clades) was only the second most likely. It has to be kept in mind, however, that these analyses were based on a reduced data set of one SNP per contig (i.e., 1,552 SNPs compared with 6,807 for the other

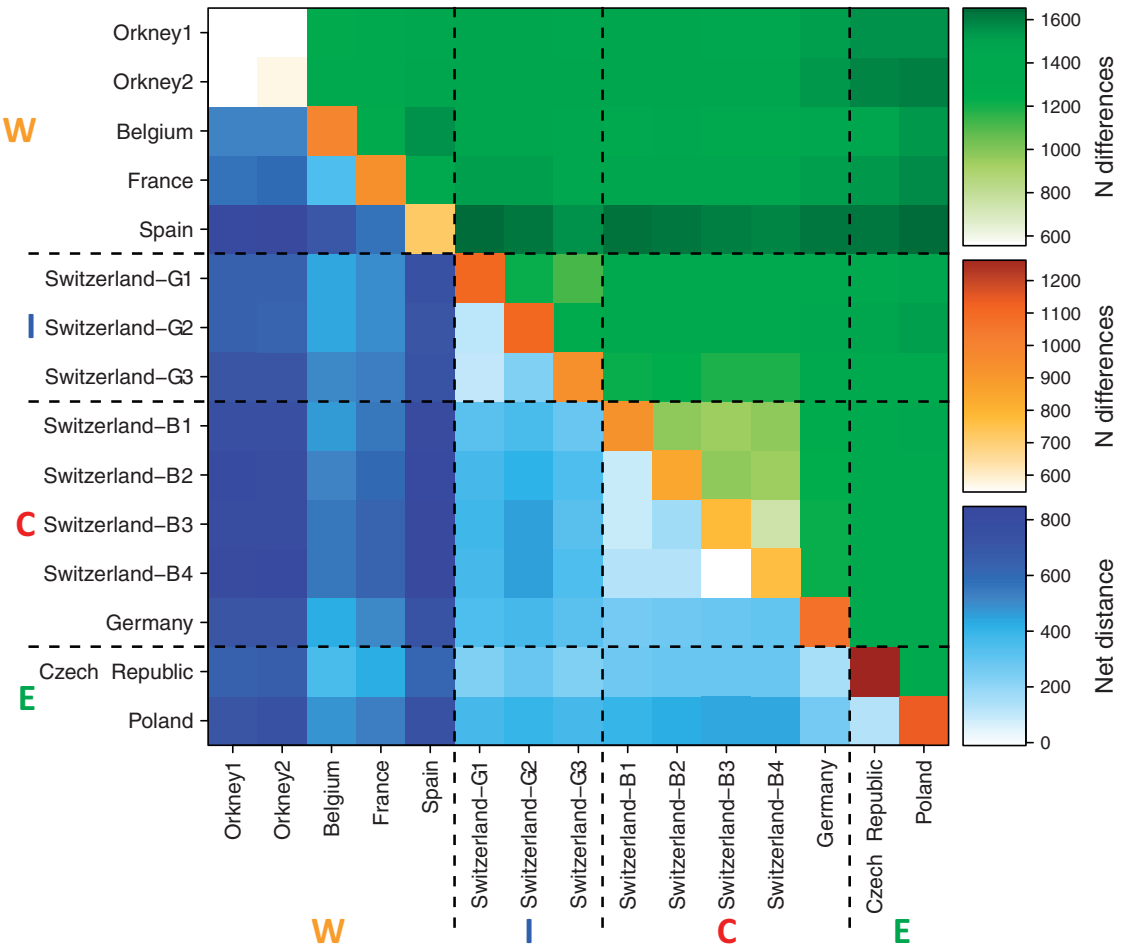

Fig. 2. Genetic variability in *Microtus arvalis*. The lower half matrix (blue) shows net average nucleotide differences between individuals at 1,552 loci. The diagonal (orange) shows the average number of nucleotide differences within individuals and the upper half matrix (green) between individuals. The dashed lines separate evolutionary lineages (W: Western; I: Italian; C: Central; E: Eastern).

Table 2. Influence of Heterozygous SNP Handling on Divergence Time Estimates between Evolutionary Lineages of *Microtus arvalis*.

Heterozygous Position Handling	No. of Trees	Divergence Time between Lineages (Years BP ± Standard Deviation)		
		W − (I, C, E)	I − (C, E)	C–E
RRHS[a]	5,000	15,306 ± 1,188	13,947 ± 1,165	13,703 ± 1,060
Average no. of differences[b]	1	15,364	13,967	13,729
Ambiguity codes	1	42,164	28,352	25,278
Totally removed	1	36,299	19,467	15,524

NOTE.—BP, before present; W, Western; I, Italian; C, Central; E, Eastern.
[a]RRHS strategy.
[b]NJ trees based on average number of nucleotide differences between individuals.

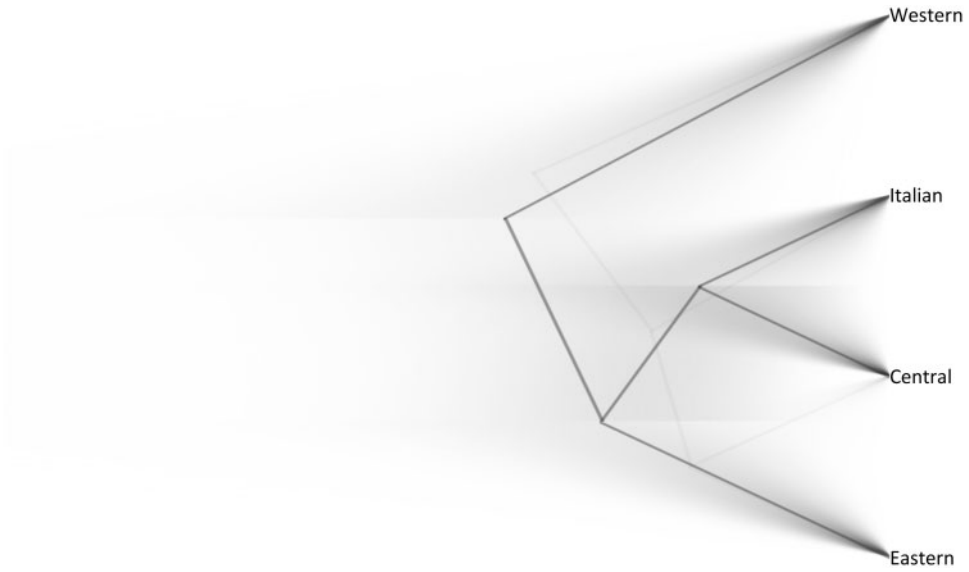

FIG. 3. Variance in phylogenomic trees estimated with SNAPP. Trees were inferred from 50 SNAPP analyses based on 1,552 randomly chosen independent SNPs each. The 15 *Microtus arvalis* individuals were grouped according to their evolutionary lineage. Trees following the most probable tree topology are drawn in dark gray and the rest in light gray. The majority consensus tree is shown with a heavy and the minority consensus tree with a lighter gray line.

methods) because SNAPP assumes SNPs to be unlinked. The species tree approach of SNAPP provides the possibility to infer divergence times directly (Bryant et al. 2012), thus in our case without the calibration with the Orkney samples. With a mutation rate estimate of 1.1×10^{-8} from mouse (Drake et al. 1998) and assuming an average of two generations per year (Hausser 1995; Hamilton et al. 2005; Martínková et al. 2013), the divergence of the Western lineage from the rest was estimated to have occurred 22.9 ± 10.6 Ka, the split of the Eastern lineage 17.1 ± 8.4 Ka, and divergence between Italian and Central 11.4 ± 5.7 Ka. These estimates were based on SNAPP analyses, which use only polymorphic loci, but if we consider that our data set contains 10.9% monomorphic loci (table 1), a correction by this factor would result in divergence time estimates of 25.7 ± 11.9 Ka for the Western, 19.2 ± 9.4 Ka for the Eastern, and 12.7 ± 6.4 Ka for the Italian and Central lineages. However, these estimates have to be viewed with much caution because the genomic mutation rate of *M. arvalis* may be different from mouse and the generation time may vary between one and three generations per year depending on environmental conditions (Ryszkowski et al. 1973; Hausser 1995; Tegelstrom and Jaarola 1998; Hamilton et al. 2005). Thus, the application of different combinations of generation times and mutation rates from mouse ($1.1–4.9 \times 10^{-8}$) (Drake et al. 1998; Ro and Rannala 2007; Lynch 2010) yields widely varying estimates for divergence times between lineages (1.7–45.8 Ka and, e.g., between 3.4 and 45.8 Ka for the separation of the Western lineages) from the uncalibrated SNAPP analyses.

Heterozygous SNP Handling

Different approaches for handling the heterozygous SNP positions had a large impact on the estimations of divergence times (table 2 and supplementary table S1, Supplementary Material online), but they did not affect the overall tree topology (fig. 4). NJ trees based on RRHS resulted in basically the same tree and divergence time estimations as NJ trees based on the average number of base differences. The phylogenomic tree based on a data set encoding heterozygous positions as IUPAC ambiguity codes showed relatively shorter end branches (fig. 4*c*). Here, two individuals with

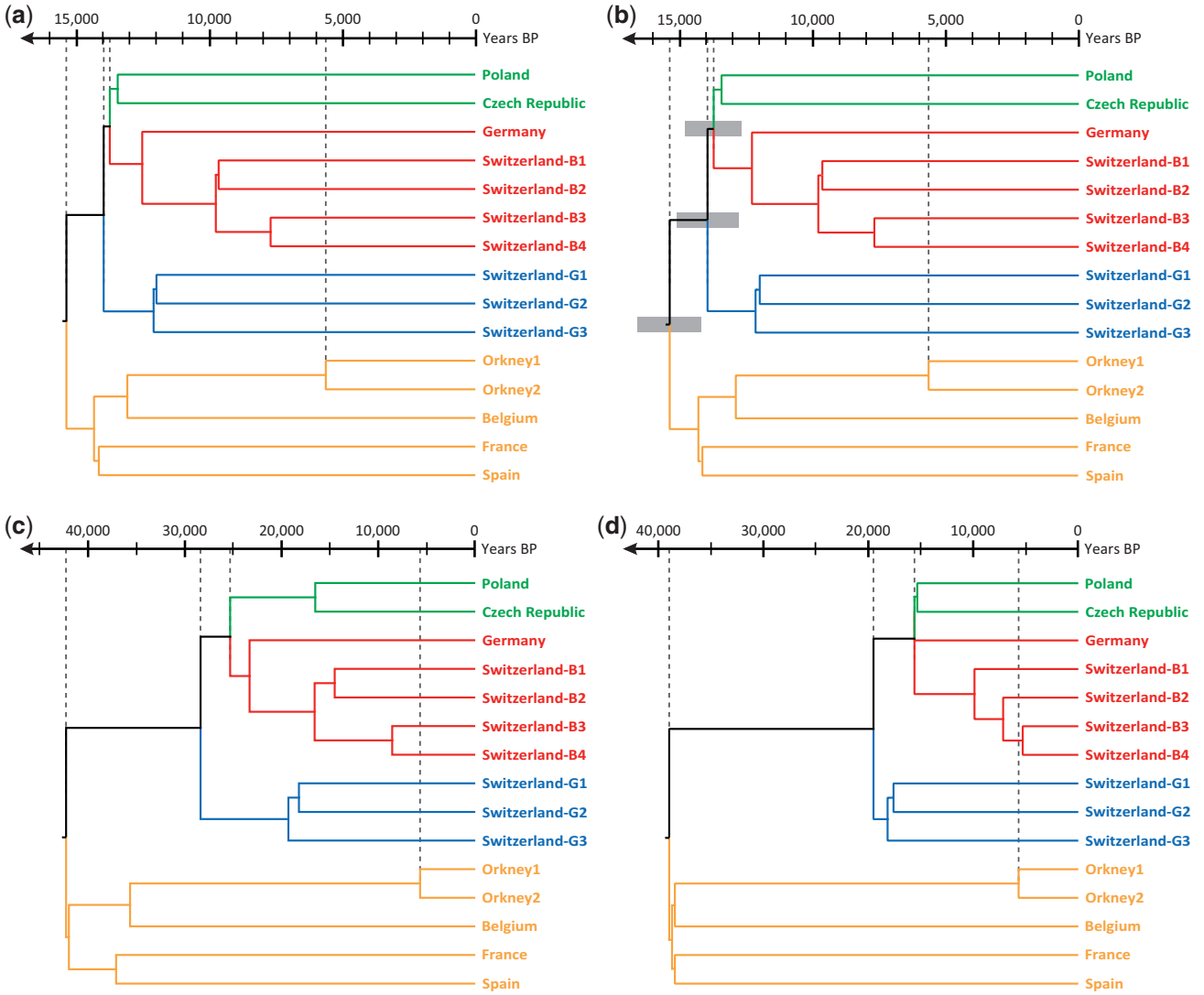

Fig. 4. Influence of heterozygous SNP handling on phylogenomic trees. Midpoint-rooted NJ trees from 1,552 AFLPseq loci of 15 *Microtus arvalis* individuals. Different heterozygous SNP handling strategies were applied: (*a*) average number of differences, (*b*) RRHS, (*c*) ambiguity codes, and (*d*) totally removed. Divergence times are calibrated on the Orkney individuals' split and are given in years before present (BP). The gray bars show the standard deviations of divergence time. The colors represent the evolutionary lineages Eastern (green), Central (red), Italian (blue), and Western (orange).

the same ambiguity code are treated as identical, whereas the mean difference between these individuals is in fact 0.5. This partial "removal" of nonfixed differences due to the use of IUPAC codes led to much older estimates of divergence times between the evolutionary lineages. The complete removal of all heterozygous SNP positions (88.4%) from the data set led to a drastic reduction of informative sites (only 789 SNP positions left) and to a bias in the relative branch lengths (fig. 4*d*). The resulting divergence time estimations were much older than with RRHS for the Western lineage and slightly older for the split of the Italian lineage and the Eastern and Central lineages.

Repeated subsampling from our data set showed that even a strong reduction of the number of loci had no effect on the tree topology (supplementary fig. S3, Supplementary Material online) and only a small effect on the divergence time estimations (table 3). The standard deviation of the divergence times increased with fewer loci, but the mean remained almost identical to the full data set of 1,552 loci. However, with very low numbers of loci (e.g., 100), the divergence time estimates were roughly 1,300 years older and the standard deviations six times larger. Additionally, the topology was less stable (low support values for internal nodes) with lower numbers of loci. An evaluation of parameters used in the processing of the sequence reads showed that an even more stringent error tolerance (0.01%) for calling SNPs led to a loss of 939 SNP positions but only to a slight increase of the divergence times estimates (table 3). Increasing the similarity threshold used to establish homology of contigs between individuals in the assembly process reduced SNP numbers slightly and led to a small increase of divergence time estimations between lineages (table 3).

Phylogenomic Modeling

We used computer simulations to examine whether our methodological approach and the time calibration based

Table 3. Influence of Assembly Parameters and Number of Loci in Genomic Data on Divergence Time Estimates between Evolutionary Lineages of *Microtus arvalis* Based on NJ Trees Using the RRHS Strategy.

Assembly Threshold (Similarity)	SNP Error Rate (%)	No. of Loci	No. of SNPs	Divergence Time between Lineages (Years BP ± Standard Deviation)		
				W − (I, C, E)	I − (C, E)	C–E
0.91	1.00	1,552	6,807	15,306 ± 1,188	13,947 ± 1,165	13,703 ± 1,060
0.91	1.00	1,000	4,387 ± 72[a]	15,370 ± 1,723	14,020 ± 1,656	13,777 ± 1,525
0.91	1.00	500	2,193 ± 68[a]	15,480 ± 2,683	14,170 ± 2,544	13,923 ± 2,375
0.91	1.00	100	438 ± 36[a]	16,758 ± 6,772	15,626 ± 6,402	15,160 ± 5,997
0.91	0.01	1,481	5,868	16,873 ± 1,402	15,396 ± 1,373	15,137 ± 1,249
0.92	1.00	1,522	6,630	16,173 ± 1,230	14,783 ± 1,202	14,402 ± 1,085
0.94	1.00	1,474	6,414	16,376 ± 1,232	15,003 ± 1,222	14,613 ± 1,100

NOTE.—BP, before present; W, Western; I, Italian; C, Central; E, Eastern.
[a]Mean ± standard deviation.

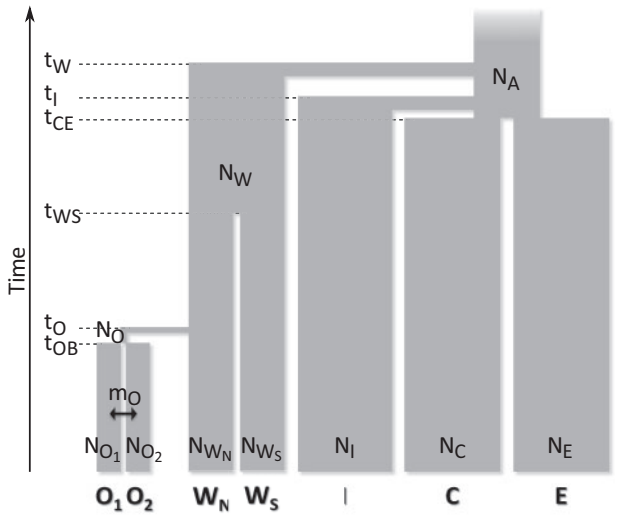

FIG. 5. Evolutionary model of the demographic history used to simulate molecular diversity in *Microtus arvalis*. The four evolutionary lineages (Western: W, Italian: I, Central: C, and Eastern: E) split consecutively from the ancestral population. The Western lineage diverged into two sublineages (W_N and W_S). Later Orkney (O) was colonized by a few individuals from W_N and split after a bottleneck into two populations (O_1 and O_2).

on the Orkney colonization is able to reproduce the evolutionary history and divergence times of *M. arvalis* realistically and to test the effects of different demographic parameters on divergence time estimations (fig. 5 and table 4). All trees based on simulated data sets mimicking our empirical data recovered the modeled evolutionary history of lineage divergence, but, depending on the model parameters, divergence time estimates were under- or overestimated (table 5). As in our empirical case, the range of divergence times estimates obtained by basing phylogenetic analyses on randomly assigned haplotypes can be rather large (supplementary table S2, Supplementary Material online), which shows the importance of informing about the variance as done by RRHS. In general, the divergence time estimates were younger if the ancestral population size was smaller (N_A = 2,000) than the population sizes of the diverged lineages (N_I, N_C, and N_E = 30,000). As expected, migration between the two Orkney populations led to an overestimation of the divergence times between the other populations. A shorter bottleneck after the colonization of Orkney decreased divergence times between lineages, as this reduces the probability that the sampled Orkney sequences coalesce during this bottleneck phase. A larger ancestral population size led to an increase in divergence time estimates, as the probability of coalescence decreases overall.

Divergence time estimations using RRHS were much less affected overall by the underlying demographic model than the estimations using IUPAC ambiguity codes or the total removal of heterozygous positions (table 6 and supplementary table S3, Supplementary Material online). Migration between the Orkney populations led generally to the strongest deviations particularly with ambiguity codes. SNAPP produced the correct topology in only 40.4% of the cases, but mutation rate-based divergence time estimations for these were relatively close to the truth except for cases with migration between Orkney populations (table 6). Scenarios involving migration are expected to affect SNAPP analyses as the underlying algorithm assumes no gene flow between groups (Bryant et al. 2012). In contrast, when divergence time estimations were based on branch lengths from the SNAPP tree and calibrated with the Orkney population split, there was almost no effect of the underlying demographic model, and divergence times were overall much less overestimated (table 6).

Discussion

The comparison of different SNP handling strategies shows that ignoring information from heterozygous sites can have a large impact on the estimation of absolute and relative branch lengths and hence bias divergence time estimations. Phylogenomic analyses of genome-wide *M. arvalis* sequence data show a clear and consistent clustering of individuals into four lineages and divergence times likely to have occurred at the end or after the LGM.

Influence of Heterozygous SNP Handling

Most phylogenetic studies ignore heterozygous sites in diploid DNA sequences, because they are not considered

relevant and/or analysis software usually does not handle diploid data. However, the number of heterozygous positions can be large as in *M. arvalis* here and in other taxa (Griffiths et al. 2000; Dehal et al. 2002; Sota and Vogler 2003; Kelleher et al. 2007; Fink et al. 2010). Excluding heterozygous positions may lead to significantly biased parameter estimates (Sota and Vogler 2003; Garrick et al. 2010), but the direction of bias may not be straightforward to infer a priori. In our analyses, calibration with the youngest split led often to overestimation of divergence times (fig. 4), but the bias may reverse for older calibration points because they may be comparatively less affected by loss of information from heterozygous sites.

Information from heterozygous positions is particularly valuable for the study of closely related or fast radiating taxa where diagnostic sites are rare. The use of phased haplotypic data is helpful but may not always be available or reliable (Browning SR and Browning BL 2011), and genomic analyses of nonmodel organisms without reference genome typically cannot take linkage disequilibrium into account. Especially in studies of multiple independent loci, phasing between loci is not possible and thus concatenation-based phylogenomic approaches may be problematic. Encoding heterozygous sites as IUPAC ambiguity codes does not solve the issue completely as it leads to an underestimation of sequence divergence between individuals and may thus bias branch length in the phylogenetic tree (Sota and Vogler 2003). Finally, choosing one of the two alleles at random can lead to incongruent and biased tree estimations (see supplementary table S2, Supplementary Material online; Weisrock et al. 2012). RRHS offers a simple solution to these

Table 4. Demographic Model Parameter Values Used in the Simulations.

Parameter		Value 1	Value 2
Population size	N_A	30,000	2,000
	N_W, N_I, N_C, N_E	30,000	
	N_{WN}, N_{WS}	15,000	
	N_O	10	
	N_{O1}, N_{O2}	3,000	
Time (years [generations][a])	t_W	15,300 (45,900)	
	t_I	14,000 (42,000)	
	t_{CE}	13,700 (41,100)	
	t_{WS}	10,000 (30,000)	
	t_O	5,600 (16,800)	
	t_{OB}	1 (3)	6.7 (20)
Migration rate (individuals per generation)	m_O	0	1

NOTE.—A, ancestral; W, Western; I, Italian; C, Central; E, Eastern; O, Orkney; W_N and W_S, Western sublineages; O_1 and O_2, Orkney populations; WS, Western lineage split; OB, Orkney bottleneck.

[a]Assuming three generations per year.

Table 5. Influence of Demographic Model Parameters on Divergence Time Estimates between Evolutionary Lineages of *Microtus arvalis* Based on NJ Trees Applying the RRHS Strategy to Simulated Genomic Data.

N_A	Orkney Migration m_O (Individuals per Generation)	Orkney Bottleneck Duration $t_O - t_{OB}$ (Generations)	Divergence Time between Lineages (Years BP ± Standard Deviation)[a]		
			W − (I, C, E) (15,300)	I − (C, E) (14,000)	C − E (13,700)
2,000	0	20	11,691 ± 903	10,788 ± 798	10,571 ± 742
2,000	0	3	7,987 ± 628	7,370 ± 555	7,211 ± 504
2,000	1	20	23,391 ± 1,828	21,540 ± 1,586	21,206 ± 1,471
2,000	1	3	18,634 ± 1,456	17,196 ± 1,275	16,842 ± 1,167
30,000	0	20	18,058 ± 1,211	17,394 ± 1,108	17,169 ± 989
30,000	0	3	10,191 ± 695	9,829 ± 608	9,684 ± 550
30,000	1	20	41,516 ± 2,853	40,028 ± 2,492	39,519 ± 2,313
30,000	1	3	27,503 ± 1,900	26,500 ± 1,667	26,144 ± 1,497

NOTE.—N_A, ancestral population size; BP, before present; W, Western; I, Italian; C, Central; E, Eastern.

[a]Simulated divergence times are shown in parentheses.

Table 6. Consequences of Demographic Model Parameters and Heterozygous SNP Handling for Divergence Time Estimates for Evolutionary Lineages of *Microtus arvalis* Based on Simulated Genomic Data.

Heterozygous Position Handling Method	Average Correct Tree Topology (%)	Average Deviation from Simulated Divergence Time			
		Overall	Orkney Migration	Shorter Orkney Bottleneck	Larger N_A
RRHS	74.4	0.33	1.33	−0.32	0.57
RRHS[a]	75.0	0.33	1.33	−0.32	0.56
Ambiguity codes	74.8	0.80	7.71	−0.30	0.86
Totally removed	53.9	0.13	2.88	−0.25	0.68
SNAPP	40.4	6.69	14.70	0.09	−0.05
SNAPP[b]	40.4	1.28	−0.07	−0.06	−0.24

[a]Only one random SNP per locus as in SNAPP analyses.

[b]Divergence time estimations based on Orkney calibration.

issues and provides information on the uncertainty in the results. With our data sets it led to essentially the same trees as a distance-based method fully integrating heterozygous information (see fig. 4*a* and *b*). The main advantage of RRHS, however, is that it can be used with any phylogenetic method or existing phylogenetic program. The need to infer phylogenetic trees many times can be limiting for time consuming tree algorithms, but this can be overcome by using speed optimized phylogenetic programs (like RAxML, Stamatakis 2006) or by running analyses in parallel on a cluster.

Robustness of Phylogenomic Divergence

Our simulations support that the clustering of *M. arvalis* individuals in four evolutionary lineages is unlikely to be an artifact of the particular analysis methods or the specific multi locus data set. In our phylogenomic analyses, we concatenated information from many loci—an approach which can lead to wrong species trees if gene trees are highly heterogeneous and thus the assumption of homogenous tree topologies across loci is violated (Kubatko and Degnan 2007). Incongruence between gene and species trees is particularly likely in situations of frequent incomplete lineage sorting as with recently diverged taxa (short branch lengths) and large population sizes (Degnan and Rosenberg 2006; Liu et al. 2010; Kumar et al. 2012). In such situations, the most probable gene tree may be mistaken for the species tree simply because the majority of nucleotide positions support it (Liu and Edwards 2009). Several species tree methods have been developed to solve these issues, but they seem currently unable to handle such a large number of loci as we obtained here with high-throughput sequencing methods, where each locus consists of a DNA sequence containing a few polymorphic sites making it difficult to correctly resolve its gene tree (Leache and Rannala 2011). Liu and Edwards (2009) showed that distance-based methods may correctly recover the species topology even when gene trees are highly heterogeneous due to the presence of deep gene lineages. The average sequence distances are often proportional to the average coalescence times of genes, which were shown to be a consistent estimate of the species tree topology (Liu and Edwards 2009).

In our case, the consistent topology among NJ, ML, and Bayesian approaches with the full data set and with scaled-down numbers of loci suggest a very strong signal of evolutionary divergence within *M. arvalis*. The analyses with the species tree program SNAPP favored a slightly different order of divergence of the evolutionary lineages than the other methods. SNAPP implements a coalescent model and thus incorporates incomplete lineage sorting, which can cause issues in other methods. Our results suggest that this may come at the expense of a need for more data, because our simulation results show that SNAPP has difficulties to identify the correct topology, whereas NJ was able to recover it with the same number of SNPs (table 6). Nevertheless, the true topology was still the second most likely for SNAPP, and further speed-optimized developments of species tree approaches may show their strengths particularly with even larger genomic data sets now at hand.

Divergence Times between Evolutionary Lineages

The estimated maximum divergence times between the four evolutionary lineages of 13.7–20.4 Ka correspond to the end of the last glaciation period (Weichselian, between 10 and 100 Ka) and are at the lower range of previous estimates based on mtDNA (several hundred Ka to 51.9–14.6 Ka; Haynes et al. 2003; Fink et al. 2004; Heckel et al. 2005; Tougard et al. 2008). Nuclear microsatellite markers provided also relatively recent divergence time estimates (24.8–6 Ka; Heckel et al. 2005) compatible with our genomic estimates. These results correspond quite well to the scenario proposed for this relatively cold-tolerant species by Heckel et al. (2005) in which the ancestors of current *M. arvalis* lineages survived the LGM not only in southern refugia located in the Balkans, Iberia, and Italy but also in ice-free steppe tundra-like areas of Central Europe (e.g., Central lineage). After postglacial expansion, the Central, Western, and Italian lineages are currently in secondary contact in the region of the Alps (fig. 1; Heckel et al. 2005; Braaker and Heckel 2009; Sutter et al. 2013). Interestingly, recent work on Central European field voles *M. agrestis* re-estimated divergence in distinct allopatric lineages also to be much more recent than before (12.6–11.4 Ka vs. several hundred thousand years; Herman and Searle 2011). It is further interesting that some of these allopatrically diverged *Microtus* lineages are already in the process of forming new species (Bastos-Silveira et al. 2012; Beysard et al. 2012; Sutter et al. 2013; Beysard and Heckel 2014), which is in line with the explosive diversification in this rodent group (Fink et al. 2010).

Allopatric separation of evolutionary lineages in *M. arvalis* is expected to result in the decoupling of demographic processes between these lineages additional to population-specific factors. For example, genomic diversity of voles from the Central lineage in Switzerland (B1–B4) is lower than in most other analyzed individuals (fig. 2). This is in agreement with population-based analyses and most likely due to the relatively recent colonization of this region of the Alps approximately 8 Ka after the melting of the glacial ice sheet (Hamilton et al. 2005; Heckel et al. 2005). The Orkney individuals also show a large reduction in genetic variability, which is most likely due to the strong bottleneck in the human-mediated colonization process of the Orkney Islands. Recent investigations of microsatellite data show reduced variability in Orkney as well and suggest that the source of the founding individuals was in the coastal area of France or Belgium (Martínková et al. 2013). In our analyses, the Orkney samples are clearly distinct but consistently closest to the sample from Belgium.

Our time calibrations are based on radiocarbon dates of the oldest *M. arvalis* bones detected during very extensive paleontological investigations of the Orkney archipelago (Martínková et al. 2013). A relatively unlikely separation of the Orkney vole populations earlier than 5.6 Ka would also push divergence between the lineages further back but ice

cover of Orkney during the last glaciation limits this to a maximum of around 12 Ka (Martínková et al. 2013), that is, a factor of two for our estimates. A more recent split between the Orkney populations would result in even younger divergence time estimates. It has been shown that divergence time estimations based on gene trees may overestimate population or species split times and thus species tree methods should be preferred (Liu and Edwards 2009), but this effect would make divergence times in our case only even more recent. Indeed, the simulations show that underestimation of the divergence time may occur depending on the evolutionary scenario and method but overestimation is more common (supplementary table S3, Supplementary Material online). Particularly, migration between the Orkney populations has a major impact on the estimates. This impact is strongest for methods explicitly assuming the absence of gene flow (e.g., SNAPP), but simple approaches like RRHS were relatively little affected by migration in the evolutionary scenarios studied here. Also the SNAPP results calibrated on the Orkney population split were little affected by the underlying demographic model.

The divergence time estimates based on ML and Bayesian methods are slightly older (17–20.4 Ka, see supplementary table S1, Supplementary Material online) than from NJ, which is probably due to partitioning and different mutation models. Inappropriate models of nucleotide evolution can have a large effect on divergence time estimations even for relatively recent events (Arbogast et al. 2002). Genome wide data sets may be especially affected by imperfect models, because very small deviations from the assumed model can lead to significantly different estimations (Kumar et al. 2012). Additionally, not taking into account rate variation among sites leads to underestimation of branch lengths (Buckley et al. 2001; Arbogast et al. 2002), and therefore, partitioned analyses do typically better than concatenated ones (Bapteste et al. 2002; Pupko et al. 2002; Delsuc et al. 2005; Rannala and Yang 2008; Kumar et al. 2012). However, the used ML and Bayesian methods assume that all genes evolved under the same tree (topology), and it is not known how branch lengths estimations are affected if this assumption is violated (Kumar et al. 2012). Our direct divergence time estimations with SNAPP are in the same range as with the other methods, but they vary widely between 1.7 and 45.8 Ka depending on the assumed mutation rate and generation time. Given the limitations in finding the correct tree topology in our simulated evolutionary scenario (table 6 and supplementary table S3, Supplementary Material online) and the sensitivity to gene flow (Bryant et al. 2012), results concerning closely related species or lineages should be evaluated very carefully.

Conclusion

High-throughput sequencing of AFLP fragments allows one to study the genome-wide evolutionary history of populations or of closely related species in nonmodel organisms as no prior knowledge of the genome is needed (van Orsouw et al. 2007; Gompert et al. 2010; Leroux et al. 2010). We were able to provide strong genomic support for four separate evolutionary lineages in the common vole and exemplified the importance of heterozygous positions for phylogenetic parameter estimates of closely related taxa. Our RRHS strategy is a simple, straightforward way to include this information in already existing analysis pipelines. Nevertheless, there is a need to integrate heterozygous site information correctly in phylogenetic analysis methods (e.g., Potts et al. 2014) as high-throughput sequencing provides access to vast genomic data sets of diploid species. Additionally, further developments of species tree approaches able to handle such large data sets and potential gene flow are desirable as those methods should be able to bridge the fields of phylogenetics and population genetics and have the potential to better resolve the evolutionary history of related species.

Materials and Methods

Samples and DNA Extraction

We analyzed 15 individuals of *M. arvalis* sampled from different populations across Europe (see fig. 1). Populations in this species are typically differentiated even at relatively small geographic scales (Heckel et al. 2005; Schweizer et al. 2007; Martínková et al. 2013). Voles were trapped with snap traps, and tissues were stored in absolute ethanol or deep frozen. Genomic DNA was extracted from tissue using a standard phenol–chloroform protocol (Sambrook et al. 1989). The quality and quantity of the DNA was determined on 0.8% agarose gels and with a spectrophotometer (NanoDrop ND-1000; Thermo Fisher Scientific, Wilmington, DE). The DNA concentration was standardized to 100 ng/μl.

AFLP Preparation

AFLP analyses were performed by adopting the protocols established by Fink et al. (2010) and Fischer et al. (2011). Six microliters of DNA (100 ng/μl) were digested for 2 h at 37 °C with 0.10 μl *EcoRI* (100 U/μl; New England Biolabs, Ipswich, MA) and 0.10 μl *MseI* (10 U/μl; New England Biolabs). In the same reaction, 1.0 μl of each adapter pair (*EcoRI* adaptors 5 pmol/μl; *MseI* 50 pmol/μl) was ligated to cutting sites by adding 0.05 μl T4 DNA ligase (2,000 U/μl; New England Biolabs), 0.55 μl BSA (1 mg/ml; New England Biolabs), 1.10 μl NaCl (0.5 M), and 1.10 μl T4 DNA ligase buffer (New England Biolabs).

Ligated samples were diluted 10 times with double distilled water (ddH_2O) before selective preamplification was performed with primers complementary to the adapters and restriction site but with an additional base (E01 = 5′-GACTG CGTACCAATTC-A-3′ and M02 = 5′-GATGAGTCCTGAGTAA C-3′). Each PCR was performed in a 40 μl volume using 8.0 μl diluted and ligated product, 1.0 μl of each primer (5 μM), 2.4 μl $MgCl_2$ (25 mM), 4.0 μl dNTP (2.5 mM), 4.0 μl 10x PCR buffer with $(NH_4)_2SO_4$ (Qiagen), and 0.32 μl *Taq* DNA polymerase (5,000 U/ml, Qiagen). The thermo cycling parameters were 72 °C for 2 min, followed by 30 cycles of denaturation at 94 °C for 20 s, annealing at 56 °C for 30 s, and extension at 72 °C for 2 min, followed by a final extension step at 60 °C for 30 min. The quality and quantity of the preamplified products were determined with a NanoDrop spectrophotometer.

Selective amplifications were performed with 21 primer combinations (supplementary table S4, Supplementary Material online). For each selective amplification, two PCRs were done in 20 µl reaction volumes each using 4.0 µl diluted (1:10 with ddH_2O) preamplification product, 0.5 µl E-primer (10 µM), 0.5 µl M-primer (10 µM), 1.2 µl $MgCl_2$ (25 mM), 4.0 µl dNTP (2.5 mM), 2.0 µl 10× PCR buffer with $(NH_4)_2SO_4$ (Qiagen), and 0.16 µl *Taq* DNA polymerase (5 U/µl, Qiagen). A touchdown PCR was performed, starting with 94 °C for 2 min, followed by 10 cycles of 94 °C for 20 s, 66 °C for 30 s, 72 °C for 2 min, where the annealing temperature dropped by 1 °C with each repetition, followed by 26 cycles of 94 °C for 20 s, 56 °C for 30 s, 72 °C for 2 min, with a final 30-min temperature hold at 60 °C. The 2× 20 µl selective amplification products were mixed together and cleaned from primers and short DNA sequences (<100 bp) with the Sigma GenElute PDR Clean-Up kit. The DNA concentrations of the PCR products were measured with a NanoDrop spectrophotometer to standardize the amount of DNA to 0.24 µg for each primer combination.

High-Throughput Sequencing and Data Processing

Five micrograms of DNA (0.24 µg DNA per primer combination) of each individual was sent to the Functional Genomics Centre Zurich (Zurich, Switzerland) and sequenced on a full run of the Roche 454 Life Science Genome Sequencer FLX Titanium. Standard molecular barcodes (Multiplex Identifiers or MIDs, 454 Life Sciences Corporation) were used to tag individual samples.

The resulting AFLP sequences were separated according to the individuals. Low-quality reads with a mean quality score below 25 were discarded (Huse et al. 2007), and the AFLP primer sequences were removed (*Eco*RI and *Mse*I restriction sites plus three selective base pairs were kept; supplementary table S4, Supplementary Material online). Reads containing both primers are fully sequenced AFLPs and were assembled with a minimum match percentage of 0.94 (van Orsouw et al. 2007). Reads containing only one primer were de novo assembled to contigs covering both primers using CLC Genomic Workbench v. 3.7 (CLC bio, Aarhus, Denmark) with a minimum match percentage of 0.94 and an overlap of 0.7. AFLP contigs resulting from the two approaches were assembled together to obtain all contigs for each individual. We will refer to these as AFLPseq for the rest of the paper. Reads without primer and nonassembled reads with only one primer were excluded from further analysis as well as contigs shorter than 50 bp. AFLPseq from all individuals were assembled (minimum match percentage of 0.91) against each other to find homologous AFLPseq. SNP calling was performed across all 15 individuals simultaneously using the UnifiedGenotyper of the Genome Analysis Toolkit (GATK) v. 1.3-17 (McKenna et al. 2010; DePristo et al. 2011). SNPs with a phred-scaled quality score less than 20 were discarded, that is, only SNPs with an error probability less than 1% were kept. In addition, we phased the SNPs within each contig with the program ReadBackedPhasing of GATK, which performs a physical phasing within a Bayesian framework based on the observed reads.

Phylogenomic Analysis

The phylogenomic analyses were based on all AFLPseq contigs, which contained at least one polymorphic position and were shared between all 15 individuals to avoid potential issues with missing data (Leroux et al. 2010; Leache and Rannala 2011). Genetic distances were calculated from concatenated AFLPseq loci with the Dnadist program of the PHYLIP package (Felsenstein 2005) using the Kimura 2-parameter distance method (Kimura 1980). Phylogenomic trees based on the genetic distances were inferred using a NJ approach implemented in the program Neighbor of PHYLIP. For display and our comparative analyses, all trees were midpoint rooted, linearized, and calibrated using MEGA v. 5.05 (Tamura et al. 2011) by applying a strict molecular clock and a maximum divergence time of 5.6 Ka between the two Orkney populations (see results).

We developed a simple new approach called RRHS to integrate the information of all alleles from heterozygous positions into phylogenomic tree estimation. In the RRHS strategy, haploid sequences for each individual are generated by randomly picking a haplotype for each of the sampled loci. If some heterozygous positions within a locus cannot be phased, the random haplotypes are picked independently for each unphased part of the locus. A Java program performing RRHS is available on our website: http://www.cmpg.iee.unibe.ch/content/softwares__services/computer_programs/rrhs/ (last accessed January 9, 2014). In a second step, the sequences are used to infer a phylogenomic tree. This is similar in spirit to Weisrock et al. (2012), but the process of random haplotype sampling and tree inference is repeated many times in RRHS, resulting in a collection of trees covering allelic and haplotypic variation widely. We used 5,000 repetitions of RRHS and summarized the resulting trees by calculating a majority rule consensus tree with mean branch lengths and standard deviations computed using the sumt command in MrBayes v. 3.1.2 (Huelsenbeck and Ronquist 2001). The shortest and longest branch lengths were obtained using the TreeAnnotator software v. 1.6.2 within the BEAST distribution (Drummond et al. 2012). Additionally, a modified blockwise bootstrap analysis of AFLPseq contigs with 5,000 repetitions was performed to infer the stability of the tree topology. In the blockwise bootstrap analysis, blocks of adjacent characters of a given length (instead of individual characters) are randomly resampled (Künsch 1989), because neighboring characters are less likely to evolve independently. For each bootstrap repetition, random haplotype sampling was applied and a NJ tree inferred. The resulting bootstrap trees were summarized by a majority rule consensus tree using the sumt command in MrBayes.

Additional ML and Bayesian trees were estimated using RAxML version 7.2.8 (Stamatakis 2006) and MrBayes version 3.1.2 (Huelsenbeck and Ronquist 2001). For both methods, a partitioned analysis was applied with an independent general-time-reversible model of substitutions and variation in

substitutions rates among sites assigned to each partition. The Bayesian analysis was run with four (three hot and one cold) simultaneous Metropolis-coupled Markov chains for 2,000,000 iterations with chain sampling every 1,000th iteration. The first 1,000 trees were discarded as burn-in based on stationarity of the log-likelihood values of the cold chain. The ML analysis was repeated for 5,000 and the Bayesian for 1,000 RRHS data sets to incorporate heterozygous information. The resulting trees of each method were summarized by calculating majority rule consensus trees with mean branch lengths and their standard deviations using MrBayes.

We also applied the species tree approach implemented in the program SNAPP (Bryant et al. 2012) to our data. SNAPP expects that SNPs are independent from each other, thus our empirical data set was reduced to one random SNP per contig (resulting in 1,552 SNPs). The individuals were grouped into the four evolutionary lineages according to the results of the NJ approaches. We applied a gamma prior for θ ($\alpha = 1$, $\beta = 200$) with independent θ on each branch. A uniform hyperprior was used for the speciation rate λ in the Yule model. SNAPP was run for 10,000,000 iterations with sampling every 2,000th iteration. The first half of the trees was discarded as burn-in. The analysis was repeated 50 times taking new random SNPs in every round, and the complete tree set was visualized with the program DensiTree (Bouckaert 2010). The trees were summarized with MrBayes by calculating majority rule consensus trees with mean branch lengths and their standard deviations. The divergence times between the lineages were estimated by multiplying the branch length with the generation time divided by the mutation rate.

Furthermore, a PCA was performed to visualize in a model-free approach the distribution of genetic variance between the *M. arvalis* individuals. The PCA was based on the SNP data transformed to an integer matrix where each individual SNP was encoded as 0 if the individual was homozygous for the first allele, 1 if it was heterozygous, or 2 if it was homozygous for the alternative allele. PCAs were calculated using the prcomp program of the stats library in R (http://www.r-project.org).

Heterozygous Position Handling

We tested the influence of different heterozygous position handling approaches on the phylogenomic tree reconstructions with two modified AFLPseq data sets. In the first one, all heterozygous SNP positions in each individual were replaced with IUPAC ambiguity codes. In the second one, heterozygous positions were completely removed in all individuals. NJ, ML, and Bayesian trees were inferred as described earlier, except that the ML and Bayesian tree analyses were repeated 100 times. An additional NJ tree based on the average number of nucleotide differences between individuals was used to evaluate the results from the three approaches (RRHS, ambiguity code, and total removal). For this purpose, two sequences per individual were constructed, one containing the first and the other the second allele. The matrix of the average number of nucleotide differences between individuals was calculated using the program MEGA. In addition, MEGA was used to calculate the average number of nucleotide differences within individuals and the net average of nucleotide differences between individuals.

Test of Data Set Properties

We also analyzed the influence of the number of loci on our phylogenomic tree reconstructions. For this, NJ trees were inferred using RRHS from reduced numbers of AFLPseq loci by randomly choosing 5,000 times each 1,000, 500, or 100 loci from the complete data set without replacement. Furthermore, the impact of some alignment parameters on the phylogenomic tree was tested. In the first approach, we increased the phred-scaled quality score threshold to 40 for accepting an SNP, which corresponds to an SNP position error probability less than 0.01%. NJ trees with the more stringent SNP acceptance rate were calculated using RRHS with 5,000 repetitions. We also tested further similarity thresholds for the alignments between AFLPseq from different individuals (92% and 94%). This higher similarity allows fewer polymorphic positions at homologous loci. SNP calling was performed with an error probability less than 1% as described earlier, and the resulting polymorphic AFLPseq contigs shared between all 15 individuals were used to infer NJ trees with RRHS (5,000 repetitions).

Phylogenomic Modeling

DNA sequences were simulated using the program fastsimcoal v. 1.1.2 (Excoffier and Foll 2011) with the evolutionary scenario for *M. arvalis* shown in figure 5 and the parameter values given in table 4. This scenario is based on mtDNA studies (Fink et al. 2004; Martínková et al. 2013) and the results of our phylogenomic analyses: the Western lineage (W) diverged first from the ancestral population, followed by the Italian lineage (I), and a split between the Central (C) and Eastern lineages (E). The Western lineage further diverged into two sublineages in the north and south of continental Western Europe (W_N and W_S). The Orkney vole populations were founded by individuals from the sublineage W_N (Fink et al. 2004; Martínková et al. 2013). After a bottleneck during introduction, the Orkney population split into two populations (O_1 and O_2) which may exchange some migrants (table 4).

The model was used to simulate 1,600 independent diploid DNA sequence loci of 125 bp length in 15 individuals mimicking our empirical data. We applied a mutation rate of 4.9×10^{-8}, which results on average in 4.4 SNP positions per locus across the 15 sampled individuals corresponding to the empirical AFLPseq data. Phylogenetic trees were reconstructed from each simulated data set by applying the RRHS strategy with 100 repetitions for which the simulation and tree inference process was repeated 100 times each. This resulted at the end in 10,000 NJ trees, which were summarized by a majority rule consensus tree with mean branch lengths and standard deviations using MrBayes as for the empirical data. The complete simulation procedure was done for each combination of the model parameters (ancestral population

size N_A, time to the Orkney bottleneck t_{OB}, Orkney migration rate m_O) resulting in eight different simulation scenarios.

Additional NJ trees were calculated from simulated data sets (100 each) with all heterozygous SNP positions either replaced with IUPAC ambiguity codes or completely removed in all individuals, or with the data set reduced to one random SNP per locus. SNAPP was also used to infer the phylogenomic trees based on one random SNP per simulated locus. The individuals were assigned to the groups used in the model (O_1, O_2, W_N, W_S, I, C, and E). Analyses were run with 2,000,000 iterations (50% burn-in) and the same priors as for the empirical data set. SNAPP analyses were repeated for 50 simulations per model. The divergence times were estimated by multiplying the branch length with the generation time divided by the mutation rate using the values of the simulations (generation time: 0.333, mutation rate of 4.9×10^{-8}) or by calibrating on the Orkney population split as for the other methods.

Supplementary Material

Supplementary material is available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The authors are grateful to the CMPG laboratory members for discussions and comments during the preparation of the manuscript. They thank Susanne Tellenbach for technical assistance, and acknowledge the helpful comments of three reviewers. This study was supported by grants from the Swiss National Science Foundation (3100A0B-126074, 31003A-127377) to L.E. and G.H.

References

Arbogast BS, Edwards SV, Wakeley J, Beerli P, Slowinski JB. 2002. Estimating divergence times from molecular data on phylogenetic and population genetic timescales. *Annu Rev Ecol Syst.* 33:707–740.

Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, Selker EU, Cresko WA, Johnson EA. 2008. Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS One* 3:e3376.

Bapteste E, Brinkmann H, Lee JA, Moore DV, Sensen CW, Gordon P, Durufle L, Gaasterland T, Lopez P, Muller M, et al. 2002. The analysis of 100 genes supports the grouping of three highly divergent amoebae: *Dictyostelium*, *Entamoeba*, and *Mastigamoeba*. *Proc Natl Acad Sci U S A.* 99:1414–1419.

Bastos-Silveira C, Santos SM, Monarca R, Mathias MD, Heckel G. 2012. Deep mitochondrial introgression and hybridization among ecologically divergent vole species. *Mol Ecol.* 21:5309–5323.

Berry RJ, Rose FEN. 1975. Islands and evolution of *Microtus arvalis* (Microtinae). *J Zool.* 177:395–409.

Beysard M, Heckel G. 2014. Structure and dynamics of hybrid zones at different stages of speciation in the common vole (*Microtus arvalis*). *Mol Ecol.* 23(3):673–687.

Beysard M, Perrin N, Jaarola M, Heckel G, Vogel P. 2012. Asymmetric and differential gene introgression at a contact zone between two highly divergent lineages of field voles (*Microtus agrestis*). *J Evol Biol.* 25: 400–408.

Bonin A, Taberlet P, Miaud C, Pompanon F. 2006. Explorative genome scan to detect candidate loci for adaptation along a gradient of altitude in the common frog (*Rana temporaria*). *Mol Biol Evol.* 23: 773–783.

Bouckaert RR. 2010. DensiTree: making sense of sets of phylogenetic trees. *Bioinformatics* 26:1372–1373.

Boussau B, Szollosi GJ, Duret L, Gouy M, Tannier E, Daubin V. 2013. Genome-scale coestimation of species and gene trees. *Genome Res.* 23:323–330.

Braaker S, Heckel G. 2009. Transalpine colonisation and partial phylogeographic erosion by dispersal in the common vole (*Microtus arvalis*). *Mol Ecol.* 18:2518–2531.

Browning SR, Browning BL. 2007. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet.* 81: 1084–1097.

Browning SR, Browning BL. 2011. Haplotype phasing: existing methods and new developments. *Nat Rev Genet.* 12:703–714.

Bryant D, Bouckaert R, Felsenstein J, Rosenberg N, RoyChoudhury A. 2012. Inferring species trees directly from biallelic genetic markers: bypassing gene trees in a full coalescent analysis. *Mol Biol Evol.* 29: 1917–1932.

Buckley TR, Simon C, Chambers GK. 2001. Exploring among-site rate variation models in a maximum likelihood framework using empirical data: effects of model assumptions on estimates of topology, branch lengths, and bootstrap support. *Syst Biol.* 50: 67–86.

Buzan EV, Forster DW, Searle JB, Krystufek B. 2010. A new cytochrome b phylogroup of the common vole (*Microtus arvalis*) endemic to the Balkans and its implications for the evolutionary history of the species. *Biol J Linn Soc.* 100:788–796.

Corbet GB. 1986. Temporal and spatial variation of dental pattern in the voles, *Microtus arvalis*, of the Orkney Islands. *J Zool.* 208: 395–402.

Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML. 2011. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nat Rev Genet.* 12:499–510.

Degnan JH, Rosenberg NA. 2006. Discordance of species trees with their most likely gene trees. *PLoS Genet.* 2:762–768.

Dehal P, Satou Y, Campbell RK, Chapman J, Degnan B, De Tomaso A, Davidson B, Di Gregorio A, Gelpke M, Goodstein DM, et al. 2002. The draft genome of *Ciona intestinalis*: insights into chordate and vertebrate origins. *Science* 298:2157–2167.

Delsuc F, Brinkmann H, Philippe H. 2005. Phylogenomics and the reconstruction of the tree of life. *Nat Rev Genet.* 6:361–375.

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet.* 43:491–498.

Drake JW, Charlesworth B, Charlesworth D, Crow JF. 1998. Rates of spontaneous mutation. *Genetics* 148:1667–1686.

Drummond AJ, Suchard MA, Xie D, Rambaut A. 2012. Bayesian Phylogenetics with BEAUti and the BEAST 1.7. *Mol Biol Evol.* 29: 1969–1973.

Edwards SV, Liu L, Pearl DK. 2007. High-resolution species trees without concatenation. *Proc Natl Acad Sci U S A.* 104:5936–5941.

Excoffier L, Foll M. 2011. fastsimcoal: a continuous-time coalescent simulator of genomic diversity under arbitrarily complex evolutionary scenarios. *Bioinformatics* 27:1332–1334.

Felsenstein J. 2005. PHYLIP (Phylogeny Inference Package) version 3.6: distributed by the author. Seattle (WA): Department of Genome Science, University of Washington.

Fink S, Excoffier L, Heckel G. 2004. Mitochondrial gene diversity in the common vole *Microtus arvalis* shaped by historical divergence and local adaptations. *Mol Ecol.* 13:3501–3514.

Fink S, Fischer MC, Excoffier L, Heckel G. 2010. Genomic scans support repetitive continental colonization events during the rapid radiation of voles (Rodentia: *Microtus*): the utility of AFLPs versus mitochondrial and nuclear sequence markers. *Syst Biol.* 59:548–572.

Fischer MC, Foll M, Excoffier L, Heckel G. 2011. Enhanced AFLP genome scans detect local adaptation in high-altitude populations of a small rodent (*Microtus arvalis*). *Mol Ecol.* 20:1450–1462.

Gaggiotti OE. 2010. Bayesian statistical treatment of the fluorescence of AFLP bands leads to accurate genetic structure inference. *Mol Ecol.* 19:4586–4588.

Garrick RC, Sunnucks P, Dyer RJ. 2010. Nuclear gene phylogeography using PHASE: dealing with unresolved genotypes, lost alleles, and systematic bias in parameter estimation. *BMC Evol Biol.* 10:118.
Gompert Z, Forister ML, Fordyce JA, Nice CC, Williamson RJ, Buerkle CA. 2010. Bayesian analysis of molecular variance in pyrosequences quantifies population genetic structure across the genome of *Lycaeides* butterflies. *Mol Ecol.* 19:2455–2473.
Griffiths AJF, Miller JH, Suzuki DT, Lewontin RC, Gelbart WM. 2000. An introduction to genetic analysis. New York: Freeman.
Hamilton G, Currat M, Ray N, Heckel G, Beaumont M, Excoffier L. 2005. Bayesian estimation of recent migration rates after a spatial expansion. *Genetics* 170:409–417.
Hausser J. 1995. Säugetiere der Schweiz: Verbreitung, Biologie, Ökologie. Basel (Switzerland): Birkhäuser Verlag.
Haynes S, Jaarola M, Searle JB. 2003. Phylogeography of the common vole (*Microtus arvalis*) with particular emphasis on the colonization of the Orkney archipelago. *Mol Ecol.* 12:951–956.
Heckel G, Burri R, Fink S, Desmet J-F, Excoffier L. 2005. Genetic structure and colonization processes in European populations of the common vole *Microtus arvalis*. *Evolution* 59:2231–2242.
Heled J, Drummond AJ. 2010. Bayesian inference of species trees from multilocus data. *Mol Biol Evol.* 27:570–580.
Herman JS, Searle JB. 2011. Post-glacial partitioning of mitochondrial genetic variation in the field vole. *Proc Biol Sci.* 278:3601–3607.
Horvath JE, Weisrock DW, Embry SL, Fiorentino I, Balhoff JP, Kappeler P, Wray GA, Willard HF, Yoder AD. 2008. Development and application of a phylogenomic toolkit: resolving the evolutionary history of Madagascar's lemurs. *Genome Res.* 18:489–499.
Huelsenbeck JP, Ronquist F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* 17:754–755.
Huse SM, Huber JA, Morrison HG, Sogin ML, Mark Welch D. 2007. Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biol.* 8: R143.
Kelleher CT, Chiu R, Shin H, Bosdet IE, Krzywinski MI, Fjell CD, Wilkin J, Yin TM, DiFazio SP, Ali J, et al. 2007. A physical map of the highly heterozygous *Populus* genome: integration with the genome sequence and genetic map and analysis of haplotype variation. *Plant J.* 50:1063–1078.
Kimura M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide-sequences. *J Mol Evol.* 16:111–120.
Kubatko LS, Carstens BC, Knowles LL. 2009. STEM: species tree estimation using maximum likelihood for gene trees under coalescence. *Bioinformatics* 25:971–973.
Kubatko LS, Degnan JH. 2007. Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Syst Biol.* 56:17–24.
Kumar S, Filipski AJ, Battistuzzi FU, Pond SLK, Tamura K. 2012. Statistics and truth in phylogenomics. *Mol Biol Evol.* 29:457–472.
Künsch HR. 1989. The jackknife and the bootstrap for general stationary observations. *Ann Stat.* 17:1217–1241.
Leache AD, Rannala B. 2011. The accuracy of species tree estimation under simulation: a comparison of methods. *Syst Biol.* 60:126–137.
Leroux S, Feve K, Vignoles F, Bouchez O, Klopp C, Noirot C, Gourichon D, Richard S, Leterrier C, Beaumont C, et al. 2010. Non PCR-amplified transcripts and AFLP fragments as reduced representations of the quail genome for 454 titanium sequencing. *BMC Res Notes.* 3:214.
Liu L. 2008. BEST: Bayesian estimation of species trees under the coalescent model. *Bioinformatics* 24:2542–2543.
Liu L, Edwards SV. 2009. Phylogenetic analysis in the anomaly zone. *Syst Biol.* 58:452–460.
Liu L, Pearl DK. 2007. Species trees from gene trees: reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. *Syst Biol.* 56:504–514.
Liu L, Yu LL, Pearl DK, Edwards SV. 2009. Estimating species phylogenies using coalescence times among sequences. *Syst Biol.* 58:468–477.
Liu LA, Yu LL, Edwards SV. 2010. A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. *BMC Evol Biol.* 10:302.
Lynch M. 2010. Evolution of the mutation rate. *Trends Genet.* 26: 345–352.
Maddison WP. 1997. Gene trees in species trees. *Syst Biol.* 46:523–536.
Martínková N, Barnett R, Cucchi T, Struchen R, Pascal M, Pascal M, Fischer MC, Higham T, Brace S, Ho SYW, et al. 2013. Divergent evolutionary processes associated with colonization of offshore islands. *Mol Ecol.* 22:5205–5220.
McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al. 2010. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20: 1297–1303.
Meredith RW, Janecka JE, Gatesy J, Ryder OA, Fisher CA, Teeling EC, Goodbla A, Eizirik E, Simao TLL, Stadler T, et al. 2011. Impacts of the cretaceous terrestrial revolution and KPg extinction on mammal diversification. *Science* 334:521–524.
Mitchell-Jones AJ, Amori G, Bogdanowicz W, Krystufek B, Reijnders PJH, Spitzenberger F, Stubbe M, Thissen JBM, Vohralik V, Zima J. 1999. The atlas of European mammals. London: Poyser.
Nichols R. 2001. Gene trees and species trees are not the same. *Trends Ecol Evol.* 16:358–364.
O'Neill EM, Schwartz R, Bullock CT, Williams JS, Shaffer HB, Aguilar-Miguel X, Parra-Olea G, Weisrock DW. 2013. Parallel tagged amplicon sequencing reveals major lineages and phylogenetic structure in the North American tiger salamander (*Ambystoma tigrinum*) species complex. *Mol Ecol.* 22:111–129.
Potts AJ, Hedderson TA, Grimm GW. 2014. Constructing phylogenies in the presence of intra-individual site polymorphisms (2ISPs) with a focus on the nuclear ribosomal cistron. *Syst Biol.* 63:1–16.
Pupko T, Huchon D, Cao Y, Okada N, Hasegawa M. 2002. Combining multiple data sets in a likelihood analysis: which models are the best? *Mol Biol Evol.* 19:2294–2307.
Rannala B, Yang ZH. 2003. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics* 164:1645–1656.
Rannala B, Yang ZH. 2008. Phylogenetic inference using whole genomes. *Annu Rev Genomics Hum Genet.* 9:217–231.
Ro S, Rannala B. 2007. Inferring somatic mutation rates using the stop-enhanced green fluorescent protein mouse. *Genetics* 177:9–16.
Ryszkowski L, Goszczynski J, Truszkowski J. 1973. Trophic relationships of the common vole in cultivated fields. *Acta Theriol.* 18:125–165.
Sambrook J, Fritsch EF, Maniatis T. 1989. Molecular cloning: a laboratory manual. New York: Cold Spring Harbor Laboratory Press.
Schweizer M, Excoffier L, Heckel G. 2007. Fine-scale genetic structure and dispersal in the common vole (*Microtus arvalis*). *Mol Ecol.* 16: 2463–2473.
Sota T, Vogler AP. 2003. Reconstructing species phylogeny of the carabid beetles *Ohomopterus* using multiple nuclear DNA sequences: heterogeneous information content and the performance of simultaneous analyses. *Mol Phylogenet Evol.* 26:139–154.
Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22:2688–2690.
Stephens M, Donnelly P. 2003. A comparison of Bayesian methods for haplotype reconstruction from population genotype data. *Am J Hum Genet.* 73:1162–1169.
Stephens M, Smith NJ, Donnelly P. 2001. A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet.* 68:978–989.
Struck TH, Paul C, Hill N, Hartmann S, Hosel C, Kube M, Lieb B, Meyer A, Tiedemann R, Purschke G, et al. 2011. Phylogenomic analyses unravel annelid evolution. *Nature* 471:95–98.
Sutter A, Beysard M, Heckel G. 2013. Sex-specific clines support incipient speciation in a common European mammal. *Heredity* 110: 398–404.
Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol.* 28:2731–2739.

Tegelstrom H, Jaarola M. 1998. Geographic localization of a contact zone between bank voles *Clethrionomys glareolus* with distinctly different mitochondrial DNA. *Acta Theriol.* 43:175–183.

Tougard C, Renvoise E, Petitjean A, Quere JP. 2008. New insight into the colonization processes of common voles: inferences from molecular and fossil evidence. *PLoS One* 3:e3532.

van Orsouw NJ, Hogers RCJ, Janssen A, Yalcin F, Snoeijers S, Verstege E, Schneiders H, van der Poel H, van Oeveren J, Verstegen H, et al. 2007. Complexity reduction of polymorphic sequences (CRoPS (TM)): a novel approach for large-scale polymorphism discovery in complex genomes. *PLoS One* 2:e1172.

von Reumont BM, Jenner RA, Wills MA, Dell'Ampio E, Pass G, Ebersberger I, Meyer B, Koenemann S, Iliffe TM, Stamatakis A, et al. 2012. Pancrustacean phylogeny in the light of new phylogenomic data: support for *Remipedia* as the possible sister group of *Hexapoda*. *Mol Biol Evol.* 29:1031–1045.

Vos P, Hogers R, Bleeker M, Reijans M, Vandelee T, Hornes M, Frijters A, Pot J, Peleman J, Kuiper M, et al. 1995. AFLP—a new technique for DNA-fingerprinting. *Nucleic Acids Res.* 23:4407–4414.

Waters JM, Rowe DL, Burridge CP, Wallis GP. 2010. Gene trees versus species trees: reassessing life-history evolution in a freshwater fish radiation. *Syst Biol.* 59:504–517.

Weisrock DW, Smith SD, Chan LM, Biebouw K, Kappeler PM, Yoder AD. 2012. Concatenation and concordance in the reconstruction of mouse lemur phylogeny: an empirical demonstration of the effect of allele sampling in phylogenetics. *Mol Biol Evol.* 29: 1615–1630.

Yalden DW. 1982. When did the mammal fauna of the British-Isles arrive. *Mammal Rev.* 12:1–57.

Yalden DW. 1999. The history of British mammals. London: Poyser.

Yang ZH, Rannala B. 2012. Molecular phylogenetics: principles and practice. *Nat Rev Genet.* 13:303–314.